TOWARDS LAST-MILE IMPUTATION WITH GENERATIVE ADVERSARIAL NETWORKS

Boaz Cogan, Noah Schaffer

Northwestern University
Interactive Audio Lab

ABSTRACT

Musical source separation is the process of isolating individual components, or ”stems” (e.g. bass guitar, vocals, drums, etc.), from a mixture of instruments. Many state-of-the-art source separation algorithms operate by learning to create a ”mask” which can be applied to an input mixture to isolate each source. Recent advances in deep source separation have led to models which can extract relatively high-quality stems from a mixture. Despite the successes of these systems, even the best separator introduces noise and artifacts to the separated data. We propose a method of using a Generative Adversarial Network based on MelGAN [7] to remove these artifacts. Our system treats the separated data as though it needs to be reconstructed during “last-mile imputation”.

1. INTRODUCTION

Music source separation has been of interest for over a decade, but has recently gained significantly more traction with advancements in deep neural networks. Many of the most popular source separation architectures such as Demucs [5], Spleeter [3], and D3Net [12] use mask inference to perform separation. Mask inference involves building a ”mask”, which can be applied element-wise to an input signal to remove all parts of a mixture that do not contain a desired source. Training a mask inference model involves comparing the result of applying a mask to a mixture to the ground-truth sources and using this as a loss function for building masks for each source.

In a perfect source separator, one can think of applying a ”binary” mask, where the frequencies that correspond to a source are entirely retained and all other frequencies are entirely removed. However, in practice, it is more effective to learn a ”soft” mask, where each element of the mask contains a value between 0 and 1 rather than a binary value. This can be problematic, as the resulting stems from applying a soft mask often contain some frequencies from other sources and may be missing frequencies that are present in the ground-truth source. This results in noise and artifacts, which can lead to unnatural sounding audio stems. Thus, the output from source separators may require further processing.

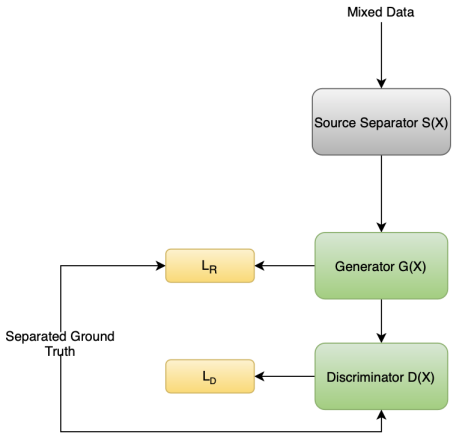

Fig. 1. Our model takes the output of a source separator and passes the result into a MelGAN, which learns to differentiate it between the ground-truth clean sources

Generative Adversarial Networks have been shown to be effective in performing imputation, which is the process of inferring missing data [13]. GANs have also been used for speech denoising, where noisy frequencies are removed from speech samples. We introduce a method that combines these two tasks to perform ”last-mile imputation.”

2. RELATED WORK

GANs have been widely used in speech denoising. Early denoising models adapted the network of an image reconstruction network, Pix2Pix [6], to perform speech denoising [10]. This model treats denoising as an image reconstruction problem in the spectrogram domain, using a pixel-wise loss between the STFTs of the clean and noisy signals.

Perceptual loss metrics such as deep feature losses [4] have emerged in recent years and have been effective in generative audio tasks as well as in speech denoising. The MelGAN architecture [7] used deep feature losses for performing conditional waveform synthesis. Leveraging MelGAN’s ability to learn features of speech, work has also been done to

translate MelGAN's architecture to the denoising domain [2].

More advanced perceptual loss metrics such as DPAM [8] and CDPAM [9] have expanded upon deep feature losses for detecting perceptual similarity between audio. HiFi-GAN [11] use these more advanced loss metrics as well as a pretrained WaveNet generator to achieve higher quality denoised speech.

Minimal work has explored using perceptual loss metrics and models like MelGAN for denoising music specifically, as most efforts are focused around speech. No known work has previously been done to post-process the results of a source separator.

3. METHODOLOGY

Many recent papers [1, 13] have shown that GANs are effective at learning to reconstruct an original image from a corrupted source. Since source separation models output suboptimal data, we can treat the separator output as the corrupted source and target the separated ground truths. As shown in Figure 1, our model only requires an existing source separator to train, allowing developers to either utilize a pretrained generator or create their own.

3.1. Pix2Pix

We first trained a Pix2Pix model which we adapted to take in audio data.The Pix2Pix generator is a UNet architecture which takes in a spectral representation of audio and returns an estimated spectorgram. The discriminator outputs a vector of confidence scores that correspond to different regions of the input spectrogram. The model was originally designed to image reconstruction, and while the system can be applied to the audio by passing in the spectrogram it can lead to undesirable artifacts. $L1$ loss is used for the pixel-wise reconstruction loss.

3.2. MelGAN

The MelGAN model is composed of a generator which takes a mel-spectrogram as input and outputs a waveform. The architecture itself is composed of a series of transposed convolutional layers, used for upsampling, followed by a residual stack of dilated convolutional layers. The main strength of the model comes from the discriminator and how the feature loss is computed.

Rather than to output a single scalar to represent whether the discriminator input was a generated image, the model outputs the model's activations after each down-sampling layer. These intermediate outputs can be considered a latent representation of the input data and are utilized when computing the feature loss, which is simply the L1 distance of these intermediate activations between the generated and the ground truth samples.

3.3. Hyperparameters

We used an Adam optimizer with a learning rate of $.0001$ and beta values of $.5$ and $.9$. We trained with a batch size of 4 over 2000 epochs.

4. EXPERIMENTS

Our experiments were evaluated purely on the model's ability to subjectively improve the quality of the input data. During evaluation we will be comparing the ground truth source, output of the source separation model, output of our model using the Pix2Pix GAN architecture, and the output of our model using the MelGAN architecture.

4.1. Dataset

For our training data we utilized a sample of the MUSDB18 train set consisting of 94 different 7 secondclips, each containing a mixture as well as separate vocals, bass, drums, and other tracks. The training data was generated by feeding the mixtures into an open source Demucs separator. Each source separated sample was split into one second segments. After separating the data and extracting the one second segments the final training set contains 2632 unique training samples.

For testing, we utilized samples from the MUSDB18 test set, which contains 50 different 7 second sound clips containing a mixture as well as separate vocals, bass, drums, and other tracks. Like the training data, we passed our testing mixtures into a Demucs separator.

4.2. Splitting Drums from Ensemble

During early experimentation with the models, we found that the drums were so distinct from all other noises that their inclusion in the training data was actually harming the overall performance of the GAN. For that reason, we have defined two experiments, training on only the separated drum data and training on the rest of the data. We test the drums and all the other sources using their respective trained models

5. RESULTS

As can be seen in Figure 2, each of the GAN architectures appear ineffective at reconstructing the ground truth when compared to the Demucs separated data. That said, each model clearly has its own strengths. The Pix2Pix model appears to be effective at identifying and enhancing the dominant frequencies in the signal, but struggles significantly to preserve low intensity regions.

Contrastingly, the MelGAN model appears to excel at correcting noisy regions of a spectrogram but is ineffective at preserving the integrity of the dominant frequencies. One possible direction for future works may be to combine the

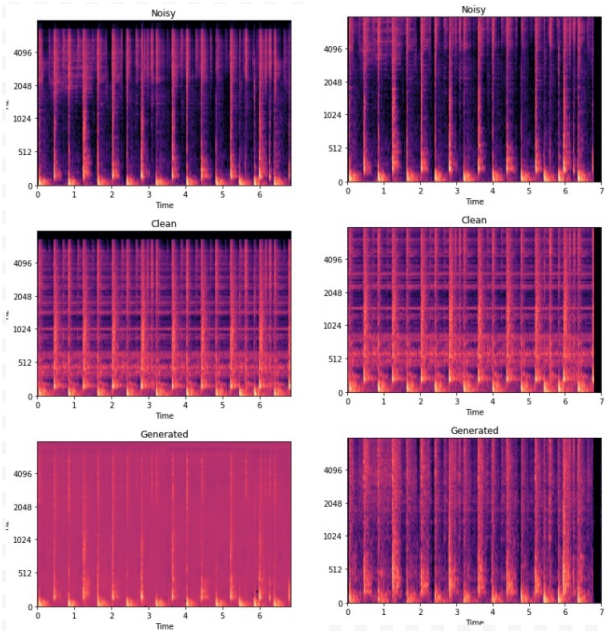

Fig. 2. Spectrograms of clean, noisy, and generated signal for a drum sample. We trained to denoise on both Pix2Pix (left) and MelGan (right)

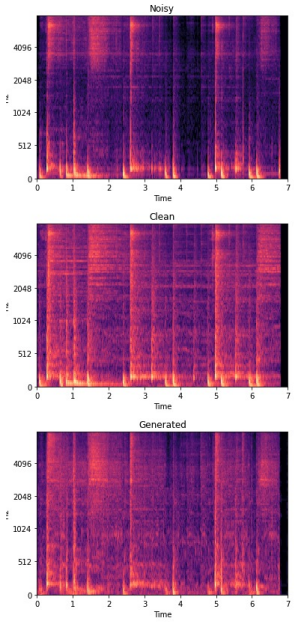

Fig. 3. The noisy, clean, and generated spectrogram for one drum sample. Our model shows it is capable of infilling missing frequencies from noisy separated drums.

two models to leverage each of their strengths, utilizing a hybrid of the deep feature and L2 distance loss to reconstruct the original signal.

We trained this model on a painfully small amount of data. Future efforts will initially be focused on expanding the dataset, using the full songs rather than seven second samples. Additionally, different source separators produce different artifacts in the separated data. By separating each mixture by a different source separator we can effectively expand the number of training samples available to our model.

It is important to note that while the generator for MelGAN is capable of reconstructing a significant amount of the original signal, it is not capable of inferring noises that were completely masked out. This can be most clearly observed by observing the periodic dots that occur between 1024 and 2048 Hz of the spectrograms in Figure 3. Since the Demucs separator has completely masked out this component of the spectrogram, our generators cannot infer any meaningful information to reconstruct it. Although MelGAN was successful at reconstructing the resonant sounds from cymbals and snares, it also introduced a static effect to the audio that significantly harms the signal quality.

6. FUTURE WORK

The weaknesses in our model may be attributed to a small or incomplete dataset. We believe that using a larger dataset could lead to significant improvements in model performance. In addition to this, most state-of-the-art source separators are trained on MUSDB18 and thus our training data may be biased in favor of the source separator. Training on more diverse data may help eliminate this bias.

Additionally, we only trained on the output of a Demucs source separator. Training on a wider variety of source separators may allow the model to better generalize artifacts across separators and can make our model more universal.

We found that when training on Pix2Pix, more complex and deep generators led to better performance. Expanding the size of the MelGAN generator and/or adapting the strucutre of the generator may be a future direction for improving model performance. Alternatively, we can train the model on a smaller receptive field. This would be similar to expanding the size of the model but may come at the cost of reduced temporal coherence.

We would also like to experiment with more complex denoising models and perceptual loss metrics. An obvious direction for this would be adapting HiFi-GAN [11], which has shown impressive results in improving the perceptual quality of noisy speech.

7. CONCLUSION

Our results have shown that when trained on the output of a Demucs, a GAN architecture is able to adequately fill missing frequencies for drums. For all other sources, we experienced generator collapse and thus were unable to obtain any meaningful results. Though the model is able to perform infill on

noisy sources, it still introduces noise and does not serve as a perceptual improvement to noisy audio from a source separator. Thus, further work must be done to ensure our model can continue to perform infill on noisy audio while not adding additional static noise. We believe our results show that with additional work, this is a promising avenue for last-mile imputation.

8. REFERENCES

[1] S. Abu Hussein, T. Tirer, and R. Giryes. Image-adaptive gan based reconstruction. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2017.

[2] L. Chkhetiani and L. Bejanidze. Se-melgan – speaker agnostic rapid speech enhancement. In *arXiv*, 2020.

[3] A. Défossez, N. Usunier, L. Bottou, and F. Bach. Spleeter: a fast and efficient music source separation tool with pre-trained models. In *Journal of Open Source Software*, 2020.

[4] F. G. Germain, Q. Chen, and V. Koltun. Speech denoising with deep feature losses. In *Interspeech*, 2019.

[5] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam. Demucs: Deep extractor for music sources with extra unlabeled data remixed. In *ArXiv*, 2019.

[6] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[7] K. Kumar, R. Kumar, T. de Boissiere, G. Lucas, W. Zhen, J. Sotelo, A. de Brebisson, Y. Bengio, and A. Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. In *Advances in Neural Information Processing Systems*, 2019.

[8] P. Manocha, A. Finkelstein, R. Zhang, N. J. Bryan, G. J. Mysore, and Z. Jin. A differentiable perceptual audio metric learned from just noticeable differences. In *Interspeech*, 2020.

[9] P. Manocha, Z. Jin, R. Zhang, and A. Finkelstein. Cdpam: Contrastive learning for perceptual audio similarity. In *ICASSP*, 2021.

[10] D. Michelsanti and Z.-H. Tan. Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification. In *arXiv*, 2017.

[11] J. Su, Z. Jin, and A. Finkelstein. Hifi-gan: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks. In *arXiv*, 2020.

[12] N. Takahashi and Y. Mitsufuji. D3net: Densely connected multidilated densenet for music source separation. In *arXiv*, 2020.

[13] J. Yoon, J. Jordon, and M. Schaar. Gain: Missing data imputation using generative adversarial nets. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.